

Efficient Computation of the Quasi Likelihood function for Discretely Observed Diffusion Processes

Lars Josef Höök^a, Erik Lindström^b

^a*Division of Scientific Computing, Department of Information Technology, Uppsala University,
Box 337, SE-751 05 Uppsala, Sweden*

^b*Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Box 118, SE-221
00 Lund, Sweden*

Abstract

We introduce a simple method for nearly simultaneous computation of all moments needed for quasi maximum likelihood estimation of parameters in discretely observed stochastic differential equations commonly seen in finance. The method proposed in this paper is not restricted to any particular dynamics of the differential equation and is virtually insensitive to the sampling interval. The key contribution of the paper is that computational complexity is sublinear in the number of observations as we compute all moments through a single operation. Furthermore, that operation can be done offline. The simulations show that the method is unbiased for all practical purposes for any sampling design, including random sampling, and that the computational cost is comparable (actually faster for moderate and large data sets) to the simple, often severely biased, Euler-Maruyama approximation.

Keywords: Quasi likelihood, Diffusion process, Conditional moment, Maximum likelihood, Stochastic differential equation

2010 MSC: 65C20, 65C30, 65C60, 68U20

1. Introduction

Most applications such as simulation or estimation involving Itô stochastic differential equations (SDEs) are in one way or another linked to the transition probabilities of the process. For example, it would be straightforward to estimate

Email address: josef.hook@it.uu.se (Lars Josef Höök)

parameters using the maximum likelihood method if transition probability density was known, but this is rarely the case in practice.

However, it is often possible to approximate the transition probability density. The probability density was obtained by brute force numerical computation of the solution to the Fokker-Planck equation (a partial differential equation) in Lo (1988); Lindström (2007) while Monte Carlo based approaches were proposed in Pedersen (1995b); Durham and Gallant (2002); Beskos et al. (2009); Lindström (2012b) and references therein. Those methods are computationally expensive, making them unsuitable for large data sets. A Gauss-Hermite series expansion of the transition probability density was proposed by Aït-Sahalia (2002), although that approach is limited to models with a specific structure.

The recent advances in collecting and storing large amounts of data are shifting the focus away from computationally slow but statistically efficient maximum likelihood methods towards computationally faster, yet not quite as statistically efficient quasi-maximum likelihood methods as the abundance of data often more than makes up for the loss of efficiency.

A simple approach based on the quasi maximum likelihood technique was introduced in Florens-Zmirou (1989) where the conditional mean and variance were obtained from an Euler-Maruyama discretization of the model, cf. Kloeden and Platen (1992). This is very efficient from a computational point of view and it was shown in Florens-Zmirou (1989) that their method is equivalent to the maximum likelihood estimator as the sampling interval goes to zero, as the bias vanishes. A higher order version of this approach is proposed in Kessler (1997). Quasi maximum likelihood methods are generally unbiased, see Sørensen (2012) provided that the mean and variance are correctly specified. The bias in the Florens-Zmirou (1989); Kessler (1997) methods therefore explicitly depends on the quality of the approximation of conditional moments, cf. Höök and Lindström (2014).

The purpose of this paper is to develop a computationally fast method quasi maximum likelihood estimator for discretely observed diffusion processes that is suitable for moderate to large data sets. We show that the computational cost is sublinear rather than superlinear due to way the moments are computed. (our simulations shows the computational complexity is comparable to that of the Euler-Maruyama scheme, and hence magnitudes faster than any approximate maximum likelihood method). This will be achieved without the bias problems associated with the Euler-Maruyama method, a property that is virtually independent of the sampling interval.

The outline of the paper is as follows. In Section 2 we formulate the statistical problem and discuss some alternative techniques for calculating conditional

moments. This is followed by Section 3 where we present a numerical implementation that results in sublinear complexity. The resulting parameter estimation algorithm is demonstrated in Section 4 on two qualitatively different diffusion processes as well as randomly sampled data followed by the conclusions being drawn in Section 5.

2. Diffusion processes and conditional moments

Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{t \geq 0})$ be a filtered probability space and let $X_t(\theta)$ be a stochastic process defined on that space that solves the following one dimensional stochastic differential equation (SDE)

$$dX_t = a_\theta(X_t)dt + b_\theta(X_t)dW_t, \quad X_{t_0} = x. \quad (1)$$

We assume throughout the paper that the drift and diffusion terms are regular enough (e.g. bounded growth and local Lipschitz, see Karatzas and Shreve (2012) for alternative conditions) to ensure existence and uniqueness of the solution.

The optimal method for estimating the parameters, θ , is the maximum likelihood estimator. Let $x_k = x(t_k)$, $k = 1, \dots, K$ be observations generated from Eq (1). The maximum likelihood estimator is defined as

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta) \quad (2)$$

where the log-likelihood function is given by

$$\ell(\theta) = \log p_\theta(x_0) + \sum_{k=1}^K \log p_\theta(x_k | x_{k-1}). \quad (3)$$

The transition probability densities, $p_\theta(x_k | x_{k-1})$ are implicitly defined by the model. The properties of the model are found by analysing the generator

$$\mathcal{L} = a_\theta(x) \frac{\partial}{\partial x} + \frac{1}{2} b_\theta^2(x) \frac{\partial^2}{\partial x^2}. \quad (4)$$

The transition probability $p_\theta(x_k | x_{k-1})$ is the solution to the Fokker-Planck equation, which is defined from the adjoint operator $(\mathcal{L}u, v) = (u, \mathcal{L}^*v)$ under the inner product (\cdot, \cdot) . The Fokker-Planck equation, when starting from x_k at time t_k and ending at t_{k+1} is given by

$$\frac{\partial}{\partial t} p_\theta(x, t) = -\mathcal{L}^* p_\theta(x, t) \quad (5)$$

with the initial condition $p_\theta(x|x_k) = \delta(x - x_k)$. This initial condition is likely to cause problems for numerical implementations of Eq. (5) due to discontinuity, see the implementation in Lo (1988) and the remedy proposed in Lindström (2007).

Another method for computing the transition probability is to use the Markov property and law of total probability, adding and integrating out an intermediate state variable, see Pedersen (1995b,a). Let $t_{k-1} < s < t_k$. It then holds that

$$\begin{aligned} p_\theta(x_k|x_{k-1}) &= \int p_\theta(x_k, x_s|x_{k-1}) dx_s \\ &= \mathbf{E}_\theta [p_\theta(x_k|x_s)|x_{k-1}] \end{aligned} \quad (6)$$

Monte Carlo methods can easily approximate that expected value, but the use of variance reduction techniques is needed for most applications, cf. Durham and Gallant (2002); Lindström (2012a).

However, we cannot expect to be able to solve either the Fokker-Planck equation Eq. (5) or the conditional expectation in Eq. (6) in closed form for more complex models. That means that the complexity of any of these approximate maximum likelihood method will be linear (in terms of expensive operations) in the number of observations.

A possible remedy are the Gauss-Hermite, see Aït-Sahalia (2002), or saddle point, see Aït-Sahalia and Yu (2006) expansions. These can be very accurate for frequently sampled data but there are also important limitations. Such as the existence of the Lamperti transform, and as well as $\Delta_k = t_k - t_{k-1}$ being small as the error typically is $\mathcal{O}(\Delta_k^{L+1})$ with L being the number of terms in the series expansion. A key operation is to employ a Lamperti transformation of the process

$$Y_t = g(X_t) \quad (7)$$

such that the dynamics of Z_t is given by

$$dY_t = f(Y_t)dt + dW_t. \quad (8)$$

The transition probability in Aït-Sahalia (2002) is given by the Hermite series approximation (here assuming that $\Delta_k = \Delta$ for all observations)

$$p_Y(y_k|y_{k-1}) = \Delta^{1/2} \phi\left(\frac{y_k - y_{k-1}}{\Delta^{1/2}}\right) \sum_{j=0}^{\infty} \eta_j^\theta H_j\left(\frac{y_k - y_{k-1}}{\Delta^{1/2}}\right) \quad (9)$$

with the coefficients given by

$$\eta_j^\theta = \frac{1}{j!} \mathbf{E}_\theta \left[H_j\left(\frac{y_k - y_{k-1}}{\Delta^{1/2}}\right) | z_{n-1} \right] \quad (10)$$

where H_j is a Hermite polynomial of order j . It is worth noting that the series expansion will not converge for all distributions and that a finite expansion may be negative for some values. The latter can be solved by considering a series expansion of the log-density, but that series may not integrate to unity. Still, the complexity is essentially sublinear as the major complexity, deriving the expansion, is performed only once.

A simpler alternative is to resort to a quasi maximum likelihood estimator, cf. Godambe and Heyde (2010); Sørensen (2012); Lindström et al. (2015). The downside is a loss of statistical efficiency as the distribution of the Maximum likelihood estimate is given by

$$\sqrt{N} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{d} N(0, I_F^{-1}), \quad (11)$$

where θ_0 is the true parameter and I_F is the Fisher information matrix defined as

$$(I_F)_{i,j} = \mathbf{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X|\theta_0) \right] \quad (12)$$

or equivalently

$$(I_F)_{i,j} = \mathbf{E} \left[\left(\frac{\partial}{\partial \theta_i} \log p(X|\theta_0) \right) \left(\frac{\partial}{\partial \theta_j} \log p(X|\theta_0) \right) \right] \quad (13)$$

whereas the distribution of the quasi maximum likelihood estimate is given by

$$\sqrt{N} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{d} N(0, J^{-1} I J^{-1}), \quad (14)$$

where

$$(J)_{i,j} = \mathbf{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \Psi(X|\theta_0) \right]$$

and

$$(I)_{i,j} = \mathbf{E} \left[\left(\frac{\partial}{\partial \theta_i} \log \Psi(X|\theta_0) \right) \left(\frac{\partial}{\partial \theta_j} \log \Psi(X|\theta_0) \right) \right]$$

with Ψ being a Gaussian density with location and scale parameters given by the conditional mean and (co)variance. That covariance is always larger or equal to the variance of the maximum likelihood estimate (this follows from the Cramer-Rao inequality), with the difference typically being rather small for nearly Gaussian models, cf. Overbeck and Rydén (1997).

2.1. Conditional moments

Parameter estimation using QML is a reason for bringing us to the topic of calculating conditional moments of the stochastic process. We will throughout this section assume that $g(\cdot)$ is a general function representing any conditional moment of interest. The conditional moment is given by

$$\mathbf{E}_\theta[g(X_k)|X_{k-1} = x_k] = \int g(x_k)p_\theta(x_k|x_{k-1})dx_k, \quad (15)$$

where $p_\theta(x_k|x_{k-1})$ is the conditional probability density.

The discussion in the previous section illustrates why Eq. (15) is intractable in the general case, which is why we have to resort to approximations. Most approximations of a conditional moment can be expressed as a weighted sum

$$\mathbf{E}_\theta[g(X_k)|X_{k-1} = x_k] \approx \sum_i \omega_i g(\xi_i). \quad (16)$$

This approximation includes techniques like Monte Carlo estimation and various deterministic quadrature rules, such as rectangular rule, the trapezoidal rule of the Gauss-Hermite quadrature.

It is also possible to approximate that conditional expectation using an Itô - Taylor expansion. Assume that the function g is $2k + 1$ times continuously differentiable. It then holds that

$$\mathbf{E}_\theta[g(X_k)|X_{k-1} = x_k] = \sum_{l=0}^L \mathcal{L}^l g(x_k) \frac{\Delta_k^l}{l!} + \mathcal{O}(\Delta_k^{L+1}), \quad \Delta_k = t_k - t_{k-1}. \quad (17)$$

Note that this expansion is not guaranteed to converge unless additional constraints are imposed on X , see Aït-Sahalia (2002) for details, but it often works quite well for small time intervals as the leading error term is $\mathcal{O}(\Delta_k^{L+1})$. This type of approximation is used compute moments in the Hermite series expansion in Aït-Sahalia (2002). It may be necessary to iterate the Itô -Taylor expansion over a series of smaller steps $\Delta t/m$ for sparsely sampled data, cf. Runge-Kutta and multistep methods, see Kloeden and Platen (1992).

An alternative, mentioned in the beginning, is to calculate conditional moments using the generator. This will require us to solve one PDEs for each moment. The Feynman-Kac (F-K) formula, see Karatzas and Shreve (2012), establishes the relation between conditional expectations and parabolic partial differential equations. Specifically, let $\tau \in [t_{k-1}, t_k]$ and define the conditional expectation

$$u(x, \tau) = \mathbf{E}_\theta[g(X_k)|X_\tau = x] \quad (18)$$

when the dynamics is given by Eq. (1). The solution to the expectation is then given as the solution to

$$\frac{\partial}{\partial \tau} u(x, \tau) = -\mathcal{L}u(x, \tau). \quad (19)$$

where the operator \mathcal{L} is defined in Eq. (4) and the initial condition is given by $u(x, t_k) = g(x)$.

There are at least two advantages of solving the adjoint problem compared to the Fokker-Planck equation. The first is that we only need to solve one single PDE for any number of observations, which should be compared to the computational complexity of the Monte Carlo and quadrature methods where it is necessary to propagate weights and/or particles between each observation. Secondly, it is more robust from a numerical point of view to solve the adjoint equation as it has a well posed initial condition (the equation is solved backwards in time) e.g. for the standard moments a polynomial, x^p .

We will later show that it is in fact enough to solve a single PDE regardless of the number of moments we are interested in. That makes the computational complexity marginal compared to that of a full blown approximate maximum likelihood estimator.

We present a conceptual summary of the pros and cons of each method respectively in Table 1. We have marked the Fokker-Planck method with a (*) since the performance of this method for small Δ_k depends strongly in the type of initial condition as described earlier.

Table 1: Feasibility of different methods for parameter estimation.

Method	Small Δ	Large Δ	Small data set	Large data set
Euler-Maruyama	Yes	-	Yes	Yes
Itô -Taylor (Eq. (17))	Yes	-	Yes	Yes
Fokker-Planck	*	Yes	Yes	-
Monte Carlo	Yes	Yes	Yes	-
Hermite series	Yes	-	Yes	Yes
Generator (F-K)	Yes	Yes	Yes	Yes

In the following section we will present a numerical approach to calculate the conditional moments from the Kolmogorov-backward equation.

3. Discretization of the Kolmogorov backward equation

Solving the Kolmogorov backward equation numerically can be achieved with a large number of different methods. We have opted for a semi-discretization with central differences in space to achieve maximum simplicity and as we will later see also the possibility to reuse calculation. The backward equation Eq. (19) is a Cauchy problem in the sense that it has a final condition defined by the conditional moment of interest, $u(x, T) = g(x)$, but lacks boundary conditions. This is similar to many option pricing problems where it is common to impose a boundary condition from asymptotic expansion of the solution, cf. von Sydow et al. (2015) for an overview of numerical techniques for computing option prices. For Eq. (19) to be well-posed it is necessary to impose boundary conditions for certain values of the coefficients. The condition when this is necessary may be found from the Fichera function (here in one dimension),

$$\mathcal{Fich} = \sum_i^{0,N} \left(a_\theta(x_i) + \frac{1}{2} \frac{\partial}{\partial x} b_\theta^2(x_i) \right) \kappa \quad (20)$$

where $\kappa = \{-1, 1\}$ are the boundary normals. Boundary conditions are not required when $\mathcal{Fich} \geq 0$, Fichera (1956). As an example the Fichera condition for the Cox-Ingersoll-Ross model at $x_0 = 0$ is given by $ab \geq \sigma^2/2$ which is also known as the Feller condition. Under the assumption that the backward equation is well-posed without boundary conditions in the sense of positive Fichera, we still need to define some conditions for the boundary values for the semi-discretized system. In Ekström et al. (2009) it was suggested to calculate the boundary values by solving a simplified backward equation at the boundaries with finite differences defined on the internal node points. We generalize this approach here by solving the full equation Eq. (19) at the boundary using interior node points. The advantage of this approach is that it does not require a large solution domain and it does not introduce a right hand side vector in the algebraic system. This enables rapid calculation of the time integration of the solution which we will utilize.

Turning to the discretization of the derivatives for the interior nodes. The first and second partial derivatives are approximated by second order central differences with $u_n = u(x_{\min} + nh)$, h being the distance between two nodes. The derivatives are then given by

$$\frac{\partial u_n}{\partial x} \approx \frac{1}{2h} (u_{n+1} - u_{n-1}) \quad (21)$$

and

$$\frac{\partial^2 u_n}{\partial x^2} \approx \frac{1}{h^2} (u_{n+1} - 2u_n + u_{n-1}). \quad (22)$$

with both approximations having errors of size $\mathcal{O}(h^2)$. Inserting the FD approximations (21 and 22) into Eq. (19) results in

$$\frac{\partial u_n}{\partial \tau} = -a_\theta(x_n) \frac{1}{2h} (u_{n+1} - u_{n-1}) - \frac{1}{2h^2} b_\theta^2(x_n) (u_{n+1} - 2u_n + u_{n-1}). \quad (23)$$

At the boundaries $0, N$ we need to solve Eq. (19) with skewed finite differences. On the lower boundary we use the following scheme:

$$\frac{\partial u_n}{\partial x} \approx -\frac{1}{h} \left(\frac{3}{2}u_0 - 2u_1 + \frac{1}{2}u_2 \right) \quad (24)$$

and

$$\frac{\partial^2 u_n}{\partial x^2} \approx \frac{1}{h^2} (2u_0 - 5u_1 + 4u_2 - u_3). \quad (25)$$

Similar approximations are used on the upper boundary,

$$\frac{\partial u_n}{\partial x} \approx \frac{1}{h} \left(\frac{3}{2}u_N - 2u_{N-1} + \frac{1}{2}u_{N-2} \right) \quad (26)$$

and

$$\frac{\partial^2 u_n}{\partial x^2} \approx \frac{1}{h^2} (2u_N - 5u_{N-1} + 4u_{N-2} - u_{N-3}). \quad (27)$$

Inserting these FD approximations (24-27) into Eq. (19) result in similarly equations as Eq. (23). Our approximation to the Kolmogorov-backward equation after approximating the spatial operator is given by,

$$\frac{\partial u_n}{\partial \tau} = \mathbf{A} u_n(\tau) \quad (28)$$

where \mathbf{A} is a banded matrix with the following elements,

$$A_{i,i+1} = \frac{a_\theta(x_i)}{2h} - \frac{b_\theta(x_i)^2}{2h^2}, \quad A_{i,i} = \frac{b_\theta(x_i)^2}{h^2}, \quad A_{i,i-1} = -\frac{a_\theta(x_i)}{2h} - \frac{b_\theta(x_i)^2}{2h^2}.$$

The first and last rows in \mathbf{A} have extra nonzero columns from the extrapolated boundary equations. Now that we have discretized the spatial operator we turn to the time discretization. In this paper we will use the matrix exponential to propagate in time, see Moler and Loan (2003). This is feasible due to the boundary

technique introduced earlier. To illustrate the benefit of our approach of including the boundary values in \mathbf{A} we consider the case when standard boundary conditions e.g. Dirichlet are used. The semi-discretized system then become

$$\frac{\partial u_n}{\partial \tau} = \tilde{\mathbf{A}} u_n(\tau) + \mathbf{b} \quad (29)$$

where \mathbf{b} contains the boundary values (here assumed to be time independent). The general solution of Eq. (29) is given by

$$u_n(t_{k-1}) = \exp\left(\tilde{\mathbf{A}}(t_{k-1} - t_k)\right) u_n(t_k) + \tilde{\mathbf{A}}^{-1} \left(\exp\left(\tilde{\mathbf{A}}(t_{k-1} - t_k)\right) - \mathbf{I} \right) \mathbf{b} \quad (30)$$

which is computationally more expensive then the solution of Eq. (28),

$$u_n(t_{k-1}) = \exp(\mathbf{A}(t_{k-1} - t_k)) u_n(t_k). \quad (31)$$

Furthermore another drawback with the classical boundary values is $\dim(\tilde{\mathbf{A}}) \gg \dim(\mathbf{A})$ since it requires a larger solution domain to avoid boundary values influencing the solution which is an additional computational cost. Returning to Eq. (31) the approximation error to the analytical solution (in terms of an exponential map) is given by,

$$u(x, t_{k-1}) = \exp(\mathcal{L}(t_{k-1} - t_k)) u(x, t_k) \quad (32)$$

$$\approx \exp(\mathbf{A}(t_{k-1} - t_k)) u_n(t_k) = u_n(t_{k-1}) + \mathcal{O}(h^2). \quad (33)$$

Since \mathbf{A} is not normal we might need to subiterate the solution in time for stability reasons. This is also required when we need to evaluate the solution at non-equidistant time intervals e.g. when the data is collected at random time instances. A subiterated solution is obtained from the following identity $\exp(\mathbf{A}\tau) = (\exp(\mathbf{A}\tau/m))^m$ where a typical good value for m can be found from the following condition $\|\mathbf{A}\tau/m\| \leq 1$ given in Moler and Loan (2003).

We use cubic spline interpolation to compute the conditional expectation for values that are not part of the finite difference grid. The interpolation error due to the cubic splines are $\mathcal{O}(h^4)$ which is dominated by the finite difference error for a dense grid.

The conditional mean and variance are accurately computed from the solution of the first moment

$$\hat{u}^{(1)}(x, t_{k-1}) = \hat{\mathbf{E}}[X_k | X_{k-1} = x]$$

and second moment

$$\hat{u}^{(2)}(x, t_{k-1}) = \hat{\mathbf{E}}[X_k^2 | X_{k-1} = x].$$

The conditional variance is then obtain through a combination of these

$$\begin{aligned}\widehat{\mathbf{Var}}[X_k|X_{k-1} = x] &= \hat{\mathbf{E}}[X_k^2|X_{k-1} = x] - \hat{\mathbf{E}}^2[X_k|X_{k-1} = x] \\ &= \hat{u}^{(2)}(x, t_{k-1}) - \hat{u}^{(1)}(x, t_{k-1})^2.\end{aligned}\quad (34)$$

3.1. Convergence

The semi discretization does not introduce any errors due to the time integration. However, the discretization of the derivatives and the interpolation do introduce errors. We can decompose the interpolated numerical solution \hat{u}^{interp} by adding and subtracting the numerical solution without interpolation \hat{u} and the true solution u . This leads to

$$\begin{aligned}\hat{u}^{interp} &= \hat{u}^{interp} \pm \hat{u} \pm u \\ &= \underbrace{\hat{u}^{interp} - \hat{u}}_{\text{Interpolation, } \mathcal{O}(h^4)} + \underbrace{\hat{u} - u}_{\text{Discretization, } \mathcal{O}(h^2)} + u \\ &= u + \mathcal{O}(h^2).\end{aligned}\quad (35)$$

Hence, the interpolation error is dominated by the discretization error if a good interpolation method is used. That trivially implies that the conditional mean can be computed with arbitrary accuracy.

Next, we find that the conditional variance is given by

$$\hat{u}^{(2),interp} - (\hat{u}^{(1),interp})^2 = (u^{(2)} + \mathcal{O}(h^2)) - (u^{(1)} + \mathcal{O}(h^2))^2 \quad (36)$$

$$= u^{(2)} - (u^{(1)})^2 + \mathcal{O}(h^2) \quad (37)$$

meaning that also the error in the conditional variance is controlled by the denseness of the finite difference grid, h . This means that both the error in the mean and covariance can be made arbitrarily small (we choose the design parameter h), leading to consistent estimates, cf. Sørensen (2012). This is in contrast to the approximate QML estimators in Florens-Zmirou (1989) and Kessler (1997) where no refinement of the estimates are possible.

The numerical quality of the method is benchmarked by comparing it against the conditional mean and variance of the Cox-Ingersoll-Ross (CIR) and the conditional mean of its inverse (iCIR). These models are defined by,

$$dX_t = a(b - X_t)dt + \sigma X_t^{1/2}dW_t \quad \textbf{CIR} \quad (38)$$

$$d\tilde{X}_t = \left[a\tilde{X}_t + (\sigma^2 - ab)\tilde{X}_t^2 \right] dt - \sigma \tilde{X}_t^{3/2}dW_t \quad \textbf{iCIR} \quad (39)$$

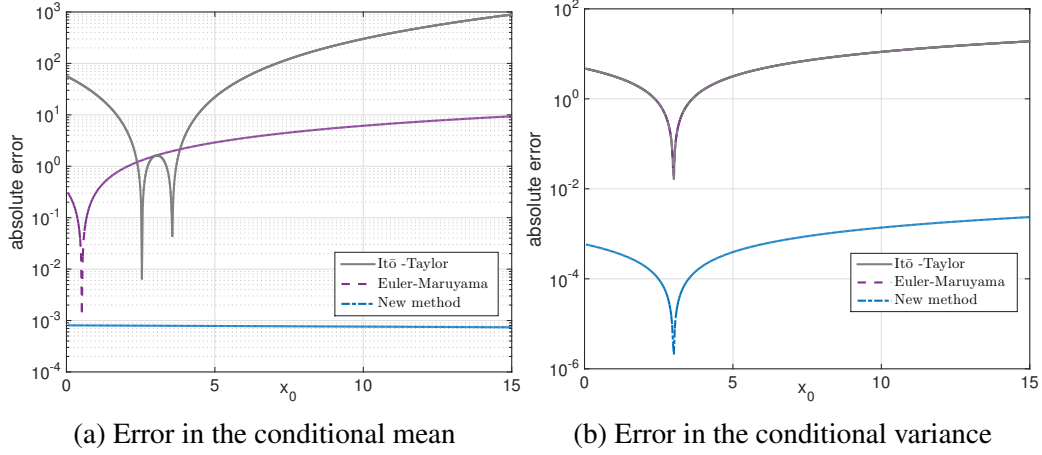


Figure 1: Absolute error for the CIR model compared between different methods for $T = 1/6$. The methods are the Euler-Maruyama scheme, truncated Itô-Taylor ($k = 1$) Eq. (17) and the proposed method. Note that the conditional var (right figure) is equal for the Euler-Maruyama and the Itô-Taylor while the mean (left figure) is not. The time T was selected such that all methods converged. The Euler-Maruyama and the generator approximation would perform even worse if larger T is used.

with $\tilde{X}_t = 1/X_t$. That means that we can (at least conceptually) compute the transition probability density as well as moments for the iCIR process.

Let $x_n \in \{x_{\min}, \dots, x_{\max}\}$ be a grid with $x_{\min} = 0.05$ and $x_{\max} = 0.15$ and $t_m \in \{0, \dots, T\}$ where the final time $T = 1/6$. The conditional mean of the iCIR process is quite lengthy and involves a Gamma function and will not be expressed here, see Ahn and Gao (1999). The absolute error between the proposed method, conditional moments using the Itô-Taylor expansion and the Euler-Maruyama scheme are compared to the exact moments for the CIR model in Figure 1.

The convergence of the numerical method is seen in Figure 2 where we have plotted the relative error (relative mean square error) as a function of spatial discretization for the CIR model using parameters $N_x = 2^{4:9} - 1$ with $h = (x_{\max} - x_{\min})/N_x$.

4. Parameter estimation

To test the applicability of the proposed framework we evaluate the quasi maximum likelihood estimator on two diffusion models. The quasi maximum likeli-

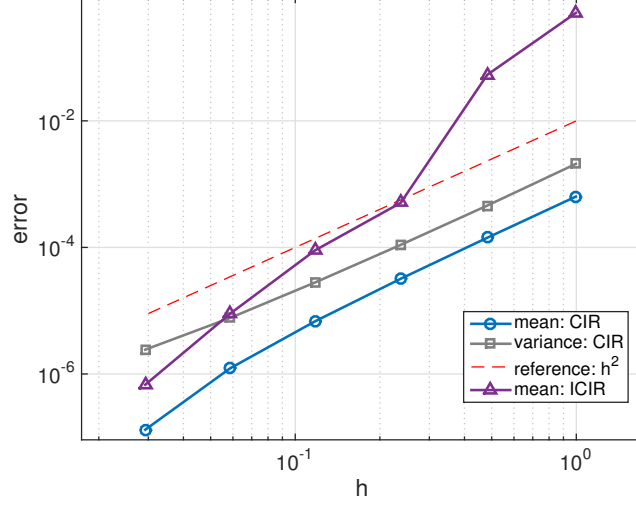


Figure 2: Spatial relative mean square error of the conditional mean for the iCIR and CIR model and conditional variance for the CIR model. Here $\tau = 1/6$ and number of time steps are $N_t = 1000$. Parameters for this test was $\{a, b, \sigma\} = \{15, 3, 2\}$.

hood estimator is defined as

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \sum_{k=1}^K \log \Psi \left(x_k; \hat{\mathbf{E}}_{\theta}[x_k | x_{k-1}], \widehat{\mathbf{Var}}_{\theta}[x_k | x_{k-1}] \right) \quad (40)$$

where $\Psi \left(x_k; \hat{\mathbf{E}}_{\theta}[x_k | x_{k-1}], \widehat{\mathbf{Var}}_{\theta}[x_k | x_{k-1}] \right)$ is the Gaussian density function with mean $\hat{\mathbf{E}}_{\theta}[x_k | x_{k-1}]$ and variance $\widehat{\mathbf{Var}}_{\theta}[x_k | x_{k-1}]$ computed with the new method. We compare the moments computed from the adjoint equation in Section 3 with the Euler-Maruyama method and to the exact moments when they are known.

4.1. Estimation on moderate data set

We also consider the inverse CIR model commonly used in interest rate modeling, see Eq. (39). This model (or actually a simplification of it) was the preferred model in for US interest rate data in the likelihood based analysis in Durham (2003). The model is challenging as the drift is non-linear but it can be shown that the conditional moments can be calculated analytically. The test data was generated from the inverse CIR model using monthly time steps $\Delta t = 1/12$. Using

$x_0 = 5$ initial value, we generated $N = 1000$ observations using the parameter $\{a, b, \sigma\} = \{15, 3, 2\}$. The first 100 observations were then discarded as burnin, leaving us with 900 observations. This was repeated 100 times in order to evaluate the estimators on independent data sets.

The estimation was conducted within the quasi maximum likelihood framework on an Intel® Core i5 @ 2.5 Ghz with 8GB of RAM. As an optimizer we used the standard Nelder-Mead, (fminsearch) in Matlab® (R2014b) with initial guess $\{10, 5, 1\}$. The results for the iCIR process is presented in Figure 3 where we see that the proposed method is unbiased whereas the Euler-Maruyama as well as the Durham-Gallant, see Durham and Gallant (2002), approximate maximum likelihood estimator are biased (the latter is due to insufficient imputation in the time domain). We also see that the proposed method is as virtually as good as the maximum likelihood estimator based on the closed form transition probability density. We also note that the Euler-Maruyama is still worse than the proposed method even when a more densely sampled data set is available for that estimator.

4.2. Estimation on randomly sampled data sets

To further test the applicability of the proposed method we estimate parameters on a simulated data set from a CIR process with samples arriving at random times, cf. Aït-Sahalia and Mykland (2003). This would be a challenging problem for a discrete time model, but can readily be handled with a continuous time model within the proposed framework. We have simulated 1000 observations from the CIR process, cf. Eq. (38) with the same parameters and burnin as for the iCIR process with random time intervals. The time intervals are uniformly distributed $t_k - t_{k-1} \sim U([1/252, 1/6])$. The results when estimating the parameters with the proposed method, Durham-Gallant method and the Euler-Maruyama is presented in Figure 4 (note that we only present the a and σ parameters as the b essentially is given by the unconditional mean of the data). It can be seen that the proposed method is unbiased even for randomly sampled data, but also that the variance for the proposed method is worse than that of the maximum likelihood estimator based on the analytical transition probability density. This is in line with our intuition as the transition density will become less Gaussian for sparsely sampled data, making the MLE the preferred estimator in that situation.

4.3. Estimation on large data sets

The challenge of large data sets is to develop fast methods. The Euler-Maruyama method is very fast (moments are given in closed form) but requires $\Delta t \rightarrow 0$ and $K\Delta t \rightarrow \infty$ for consistency, cf. Sørensen (2012). The proposed method computes

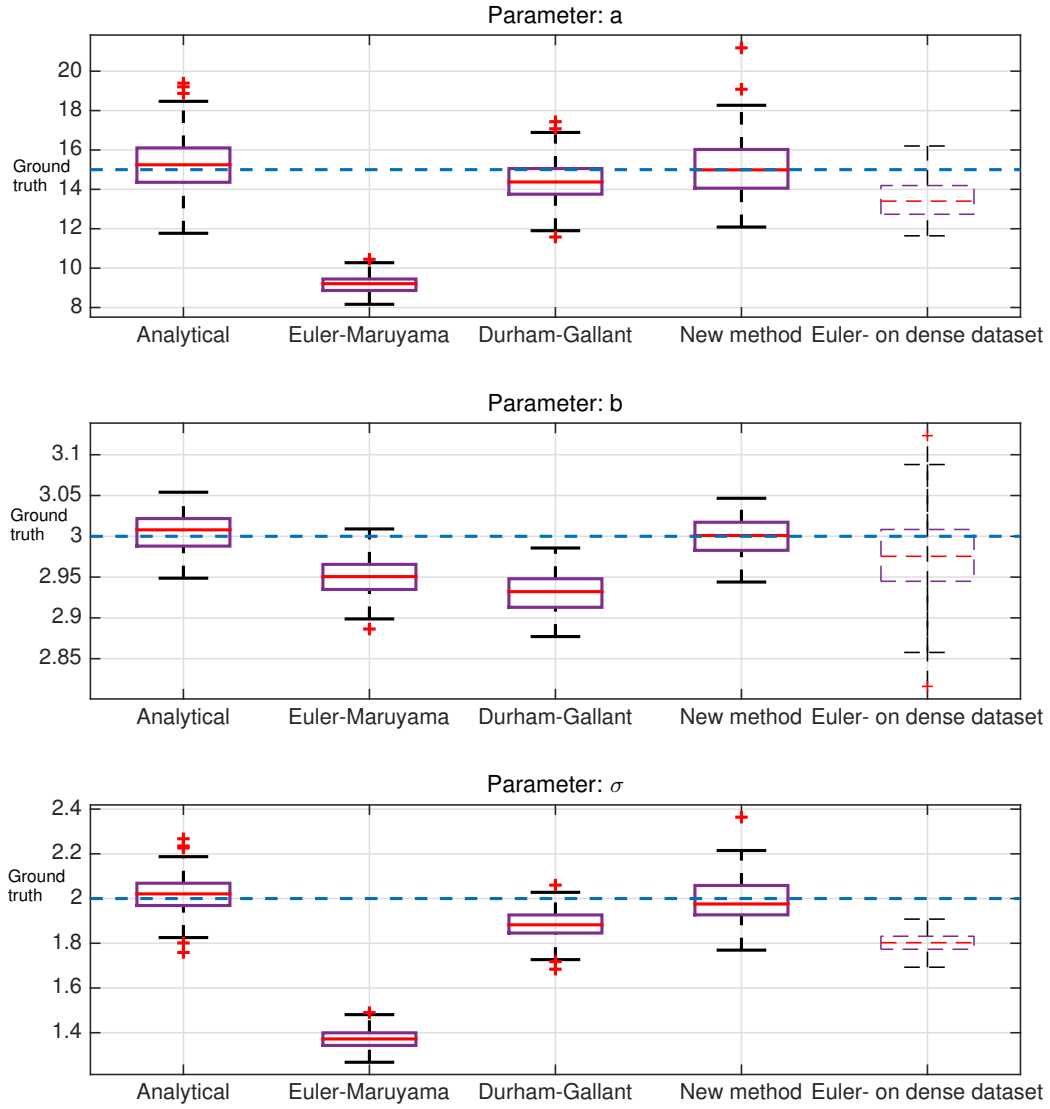


Figure 3: Estimated parameters for the iCIR model using an analytical expression for the likelihood function, the Euler Maruyama approximation, the Durham-Gallant method, our proposed method and an Euler Maruyama estimator using more densely sampled data. The data sampling interval was $\Delta t = 1/12$ for the iCIR model. Since this time step resulted in a large bias in the Euler-Maruyama method; We also ran it on a data set with $\Delta t = 1/52$, dashed boxes in the Figure. For this dense data set the Euler-Maruyama method performed better, but still suffers from bias.

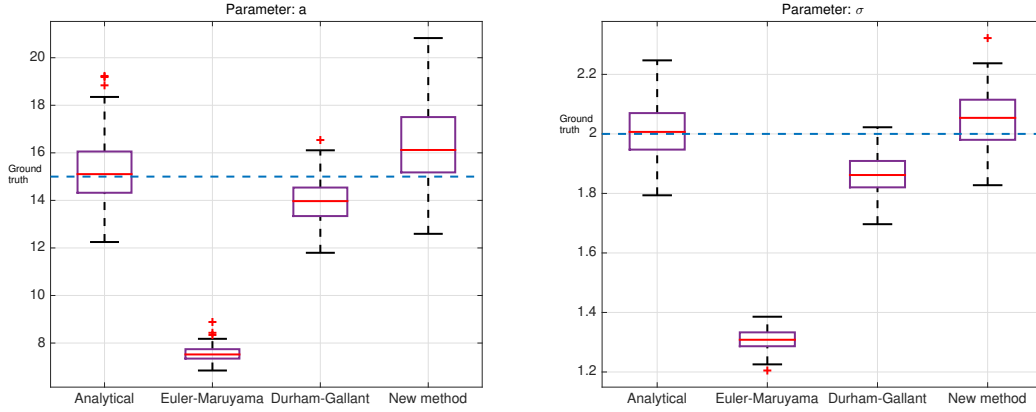


Figure 4: Estimated parameters for the CIR model with randomly sampled data using an analytical expression for the likelihood function, the Euler Maruyama approximation, the Durham-Gallant method and our proposed method. The data sampling interval was randomly distributed as $\Delta t \sim U([1/252, 1/6])$.

practically unbiased estimates, cf. Figures 3 and 4, for any sampling interval when the finite difference grid is dense enough, and will therefore only require $K \rightarrow \infty$ for consistency.

Here we evaluate the computational performance when computing the quasi likelihood function when the data consists of $K = 2\,000\,000$ observations, which makes the data set computationally infeasible for most other estimators. The parameters and sampling is the same as in Section 4.1. We present the time needed to compute the quasi likelihood function for an increasing set of observations (each point in the graph represents 1000 additional observations) for the Euler-Maruyama and the proposed method in Figure 5. The plot is based on the average taken over three simulations. The proposed method is initially more expensive than the Euler-Maruyama as we need to compute one matrix exponential to obtain the moments. However, the cost after the initial computation scales similarly as for Euler-Maruyama (it is actually somewhat cheaper for many observations), in spite that our method is consistent while the Euler-Maruyama is severely biased. This result is very encouraging as we do not need to have frequently sampled data for consistency meaning that we can work with data sets sampled over longer time horizons at the same computational cost, meaning that we could estimate certain (typically drift parameters) much better than what would be possible using only the Euler-Maruyama or similar algorithms.

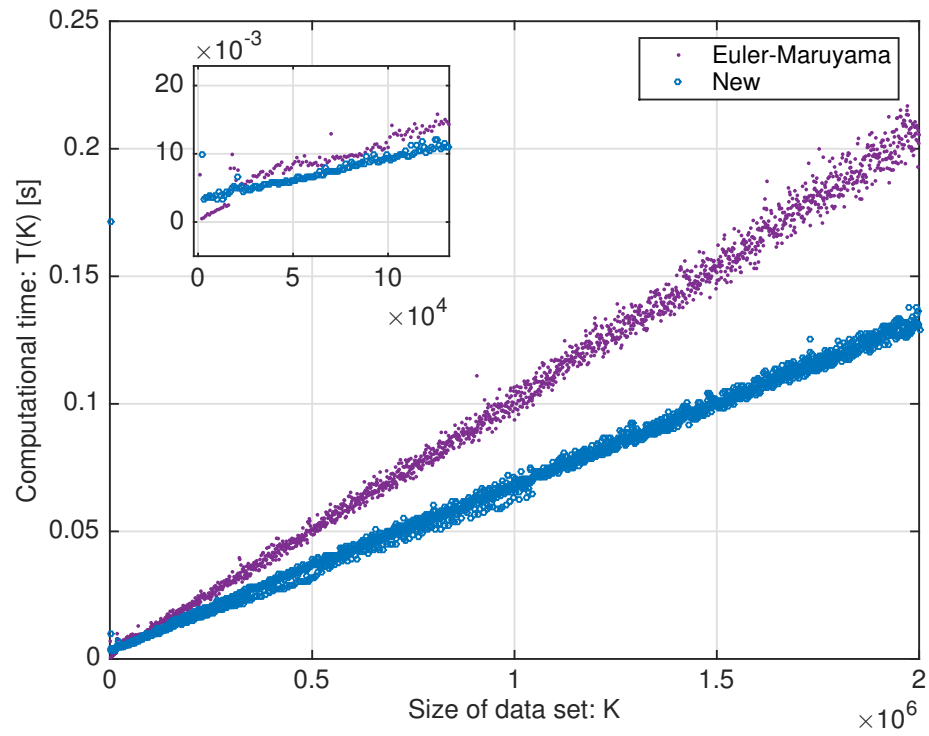


Figure 5: Computational time (wall-clock time) averaged over three consecutive measurements as a function of the data set size.

5. Conclusions

This paper introduces a framework for computing conditional moments for diffusion models based on numerical computation of the Kolmogorov-backward equation which is the adjoint to the Fokker-Planck equation with exact integration in the time domain. The numerical solution is very accurate compared to standard methods for computing conditional moments. We used the computed moments in this paper to form a quasi maximum likelihood function for parameter estimation. The method is computationally very fast, as the complexity is sublinear. All that is needed is to compute a single matrix exponential to compute all moments, regardless of the number of observations. This makes the method well suited for parameter estimation of large data sets, which was confirmed by Figure 5, in stark contrast to many approximate maximum likelihood methods for which the computational complexity typically is superlinear in the number of observations.

References

References

- Ahn, D.-H., Gao, B., 1999. A parametric nonlinear model of term structure dynamics. *Review of Financial Studies* 12 (4), 721–762.
- Aït-Sahalia, Y., 2002. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica* 70 (1), 223–262.
- Aït-Sahalia, Y., Mykland, P. A., 2003. The effects of random and discrete sampling when estimating continuous-time diffusions. *Econometrica* 71 (2), 483–549.
- Aït-Sahalia, Y., Yu, J., 2006. Saddlepoint approximations for continuous-time Markov processes. *Journal of Econometrics* 134 (2), 507–551.
- Beskos, A., Papaspiliopoulos, O., Roberts, G., 2009. Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics*, 223–245.
- Durham, G. B., 2003. Likelihood-based specification analysis of continuous-time models of the short-term interest rate. *Journal of Financial Economics* 70 (3), 463–487.

- Durham, G. B., Gallant, A. R., 2002. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics* 20 (3), 297–338.
- Ekström, E., Lötstedt, P., Tysk, J., 2009. Boundary Values and Finite Difference Methods for the Single Factor Term Structure Equation. *Applied Mathematical Finance* 16 (3), 253–259.
- Fichera, G., 1956. Sulle equazioni differenziali lineari ellittico-paraboliche del secondo ordine. *Atti Accad. Naz. Lincei. Mem. Cl. Sci. Fis. Mat. Nat. Sez. I. VIII, Ser. 5*, 3–30.
- Florens-Zmirou, D., 1989. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics: A Journal of Theoretical and Applied Statistics* 20 (4), 547–557.
- Godambe, V. P., Heyde, C. C., 2010. Quasi-likelihood and optimal estimation. In: *Selected Works of CC Heyde*. Springer, pp. 386–399.
- Höök, J., Lindström, E., 2014. A fast adjoint-based quasi-likelihood parameter estimation method for diffusion processes. Conference abstract for the Bachelier meeting.
- Karatzas, I., Shreve, S., 2012. *Brownian motion and stochastic calculus*. Vol. 113. Springer Science & Business Media.
- Kessler, M., 1997. Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics* 24 (2), 211–229.
- Kloeden, P. E., Platen, E., 1992. *Numerical solution of stochastic differential equations*. Vol. 23. Springer Science & Business Media.
- Lindström, E., 2007. Estimating parameters in diffusion processes using an approximate maximum likelihood approach. *Annals of Operations Research* 151 (1), 269–288.
- Lindström, E., 2012a. A Monte Carlo EM algorithm for discretely observed Diffusions, Jump-diffusions and Lévy-driven Stochastic Differential Equations. *International Journal of Mathematical Models and Methods in Applied Sciences* 6 (5).

- Lindström, E., 2012b. A regularized bridge sampler for sparsely sampled diffusions. *Statistics and Computing* 22 (2), 615–623.
- Lindström, E., Madsen, H., Nielsen, J. N., 2015. *Statistics for Finance*. CRC Press.
- Lo, A. W., 1988. Maximum likelihood estimation of generalized Itô processes with discretely sampled data. *Econometric Theory* 4 (02), 231–247.
- Moler, C., Loan, C. V., 2003. Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review* 45 (1), 3–49.
- Overbeck, L., Rydén, T., 1997. Estimation in the Cox-Ingersoll-Ross model. *Econometric Theory* 13 (03), 430–461.
- Pedersen, A. R., 1995a. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, 257–279.
- Pedersen, A. R., 1995b. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 55–71.
- Sørensen, M., 2012. Estimating functions for diffusion-type processes. In: *Statistical methods for stochastic differential equations*. CRC Press - Chapman and Hall, pp. 1–107.
- von Sydow, L., Höök, L. J., Larsson, E., Lindström, E., Milovanović, S., Persson, J., Shcherbakov, V., Shpolyanskiy, Y., Sirén, S., Toivanen, J., Waldén, J., Wiktorsson, M., Levesley, J., Li, J., Oosterlee, C. W., Ruijter, M. J., Toropov, A., Zhao, Y., 2015. BENCHOP – The BENCHmarking project in Option Pricing. *International Journal of Computer Mathematics*.